Weekly Report

2018.1112-2018.1118

1.This Week

Deep Learning Power Grid Project:
1.We explore the original dataset to check if the dataset itself is able to be separated. Since we discovered that it is difficult to use existing similarity measures to tell something, we made the further discoveries:

- Euclidean distance is more interpretable than other distance measures (DTW etc.) when the dimensions is very high. And it is also more efficient. This is also proved by exsing KDD papers.
- With a proper distance measure (For example, ED), it is very slow to perform the clustering algorithm (Kmeans, SOM, CURE, BIRCH, hierachical clustering) since we need to compute the distance between every two objects. So we try a more efficient clustering method **(paper 2)**.
- We are trying the cognitive similarity measure **(paper 5)** to check its accurracy and efficiency comparing to other methods. This method change the original data into image representation and compares the difference between shapes.

Power flow Project
1.Organize the materials of the previous project.
2.Make clear the tasks of the project:

- T1.Tell wether a sample is convergent.
- T2.Present the intermediate result of the equation iterations.
- T3.Discover which kind of initial conditions will lead to unconvergent results.

3.Difficulties of this project:

- Needs the intermediate results of the equation solving process.

Working Hour: (except nap and eat time)
8-9 hours / week day
8 hours on sunday
Total Working Hour this week: 51 hours.

Other
1.Write the blog for the group meeting representation.
2.Write the outline for Zongzhuang's paper.

Paper Reading
1.Finding Clusters of Different Sizes, Shapes, and Densities in Noisy, High Dimensional Data

This paper presents a clustering technique that finds clusters in high dimensional data of widely differing shapes, sizes, and densities, and especially the data of noise and outliers. It first finds the nearest neighbors of each data point and then redefines the similarity between pairs of points in terms of how many nearest neighbors the two points share. It identifies core points and then builds clusters around the core points. (The use of a shared nearest neighbor definition of similarity alleviates problems with varying densities and high dimensionality, while the use of core points handles problems with shape and size.)

2.Efficient Clustering of High-Dimensional Data Sets with Application to Reference Matching

This paper presents a clustering technique that targets at large datasets (for example, millions of data points having thousands of dimensions with thousands of clusters.).It uses a cheap, approximate distance measure to divide the data into overlapping subsets (canopies). Then it measures the exact distances only between points that occur in a common canopy. The clustering accuracy is ensured when for every traditional cluster, there exists a canopy such that all elements of the cluster are in the canopy. This method is proved to be 25% faster than existing methods.

3.iSAX: disk-aware mining and indexing of massive time series datasets

This paper improves the original SAX representation for time series by using binary encoding to get the alphabetic representation of the original SAX representation. There are three advantages for doing that. First, it supports multi-resolution representation because the representation can be easily switched between different resolutions. Second, a simple tree-based index structure can be constructed to support fast exact search. Third, it reduces computation cost.

4.iSAX 2.0: Indexing and Mining One Billion Time Series

This paper proves that indexing in massive collections of time series, building the index is most time consuming. It introduces a bulk loading mechanism to build a time series index and reduces the time cost.

5.Shape Similarity Measure Based on Correspondence of Visual Parts

This paper presents a cognitively motivated similarity measure. It can be used to retrieve similar objects in image databases (2D). The shapes are simplified by digital curve evolution to reduce influence of digitization noise and segmentation errors. It first establishes the best possible correspondence of visual parts. Then, the similarity between corresponding parts is computed and aggregated.

2.Progress

Work	Deadline	Progress
Power grid paper with Deeping learning	12.15	1.explore the original data space.
SQC Paper	-	1.Delayed